

# Data Services: Making It Happen

Heather Coates, Stacy Konkiel, Michael Witt

ACRL 2013: April 12, 2013



#dataservices





# Agenda

- Panel Introductions
- Back to the 80s
- The data challenge
- Data Services: In 6 Acts
- Discussion

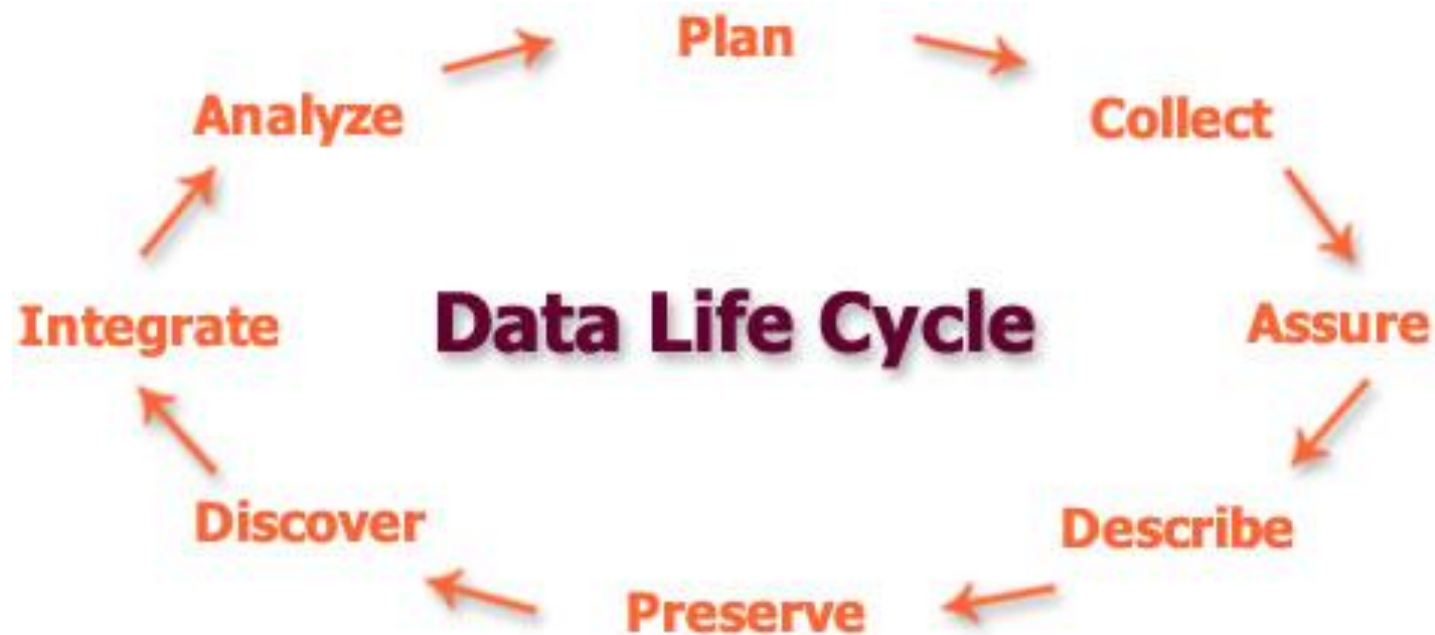
# The data challenge

- The data deluge: Volume, Velocity, Variety
- Privacy & security concerns
- Digital data vs. analog data
- Making sense of it: discovery, access, & reuse
- The case for Open Data
- The Academy responds

# Why librarians should get involved

- Trust
- Interdisciplinary/Collaborative
- Existing Infrastructure
  - Preservation
  - Digital Content

# Data Life Cycle

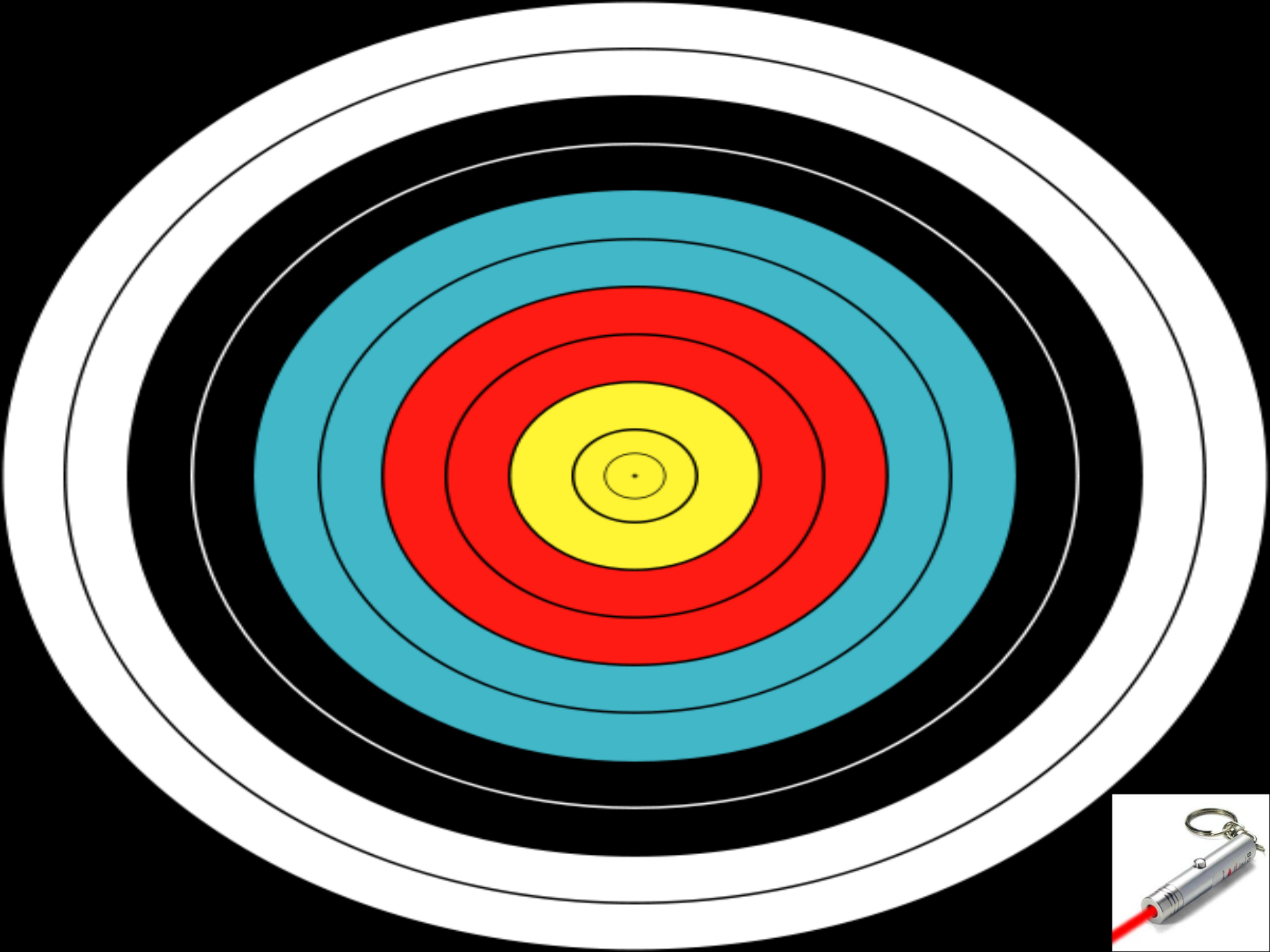


Source: <http://www.dataone.org/best-practices>

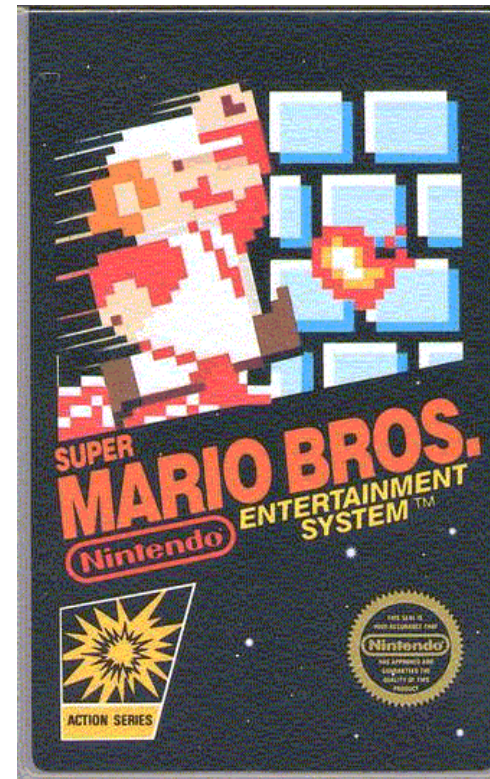
*It's about supporting research, at all phases of the life cycle*

# **Data Services: In 6 Parts**

- Consultations [Heather]
- Training [Heather]
- Metadata & Documentation [Stacy]
- Preservation [Stacy]
- Institutional Repositories [Mike]
- Data Citation [Mike]

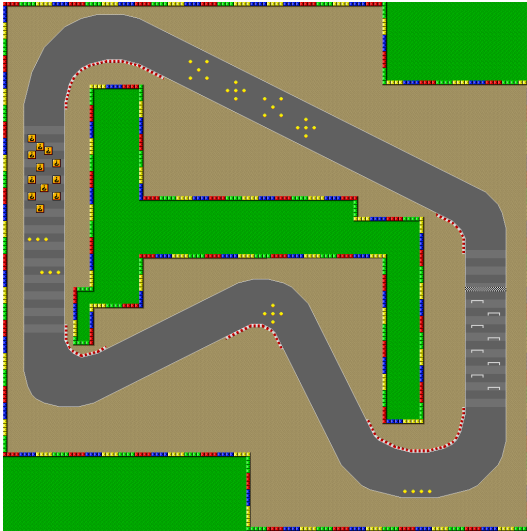




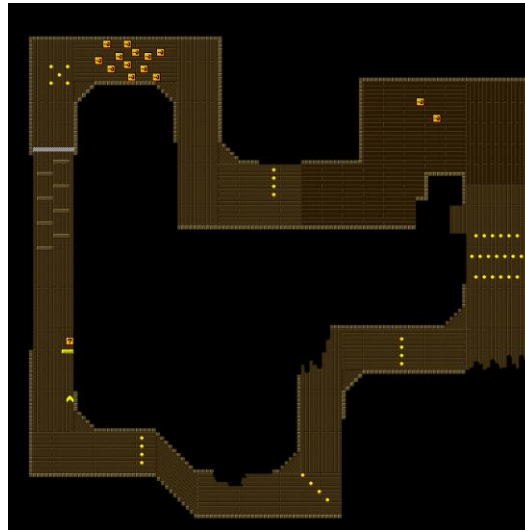




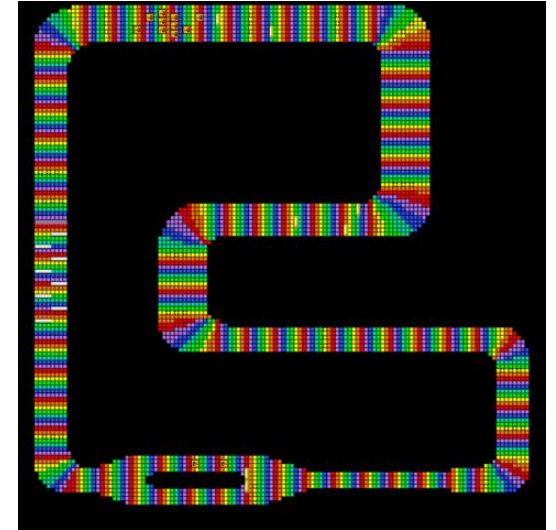
# Data Services Routes



**Mario  
Circuit 1**

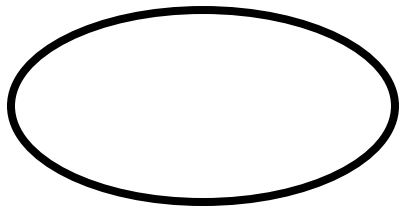
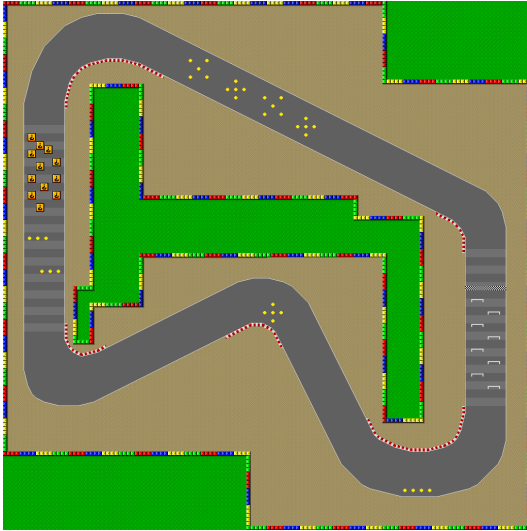


**Ghost  
Valley 2**

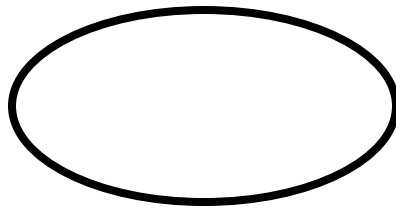
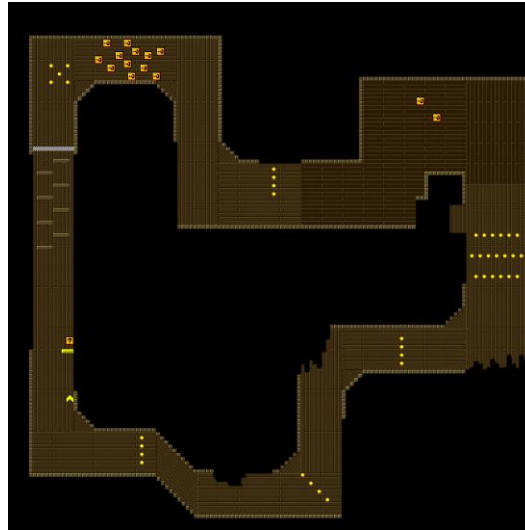


**Rainbow  
Road**

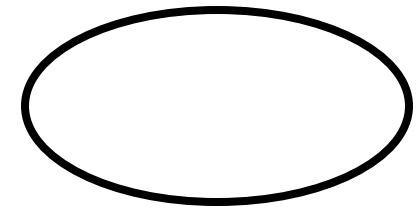
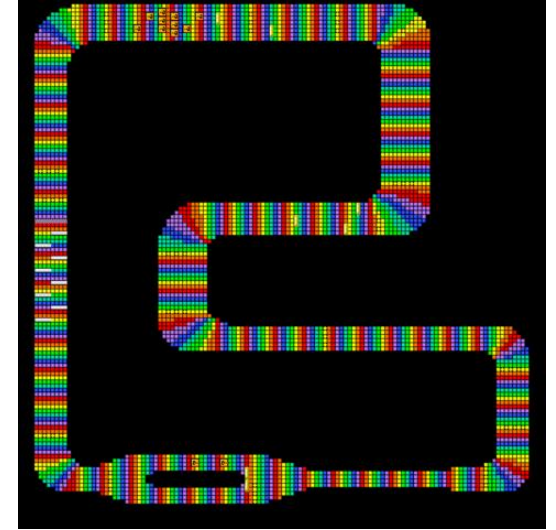
# How adventurous do you feel?



**Mario Circuit 1:**  
Well-tread, safe,  
smooth



**Ghost Valley 2:**  
Bumpy, guard  
rails



**Rainbow Road:**  
Sharp turns, no  
rails, hard to see

Data management  
plans are great!

Ugh, more  
administrative  
overhead...



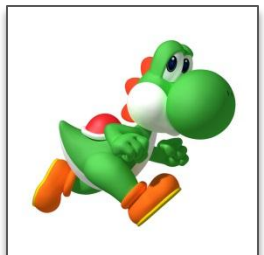
# Consulting - Be like Yoshi

- Help overcome obstacles
- Navigate unfamiliar & difficult territory
- Be a good sidekick...but don't let Mario sacrifice you to win!



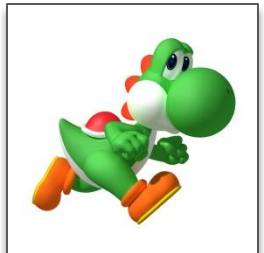
# What happens during a consult?

- Meeting the funding agency requirement
  - Gather info for draft or review and refine DMP
- Culture change
  - *Nudge* towards institutional cyberinfrastructure
  - *Suggest* better data management practices
  - *Encourage* use of IR for dissemination
  - Navigating the system - referral to other units



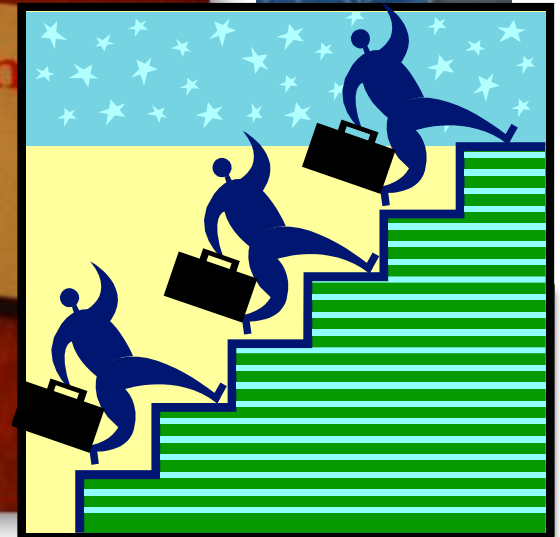
# Data management plans

- Describing data
- Defining standards
- Intellectual property and rights management
- Licensing
- Archiving and preservation



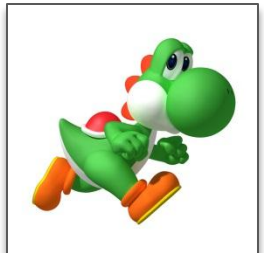


# Strategies for success

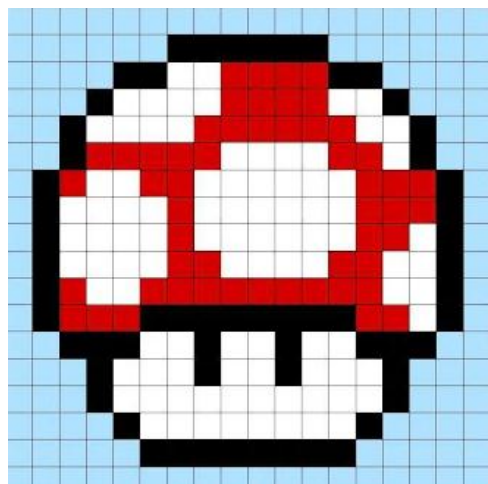
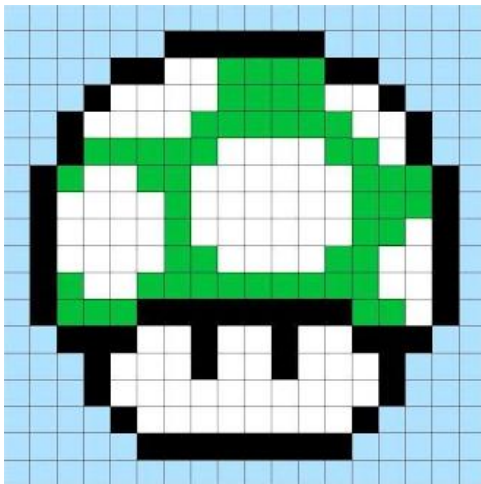
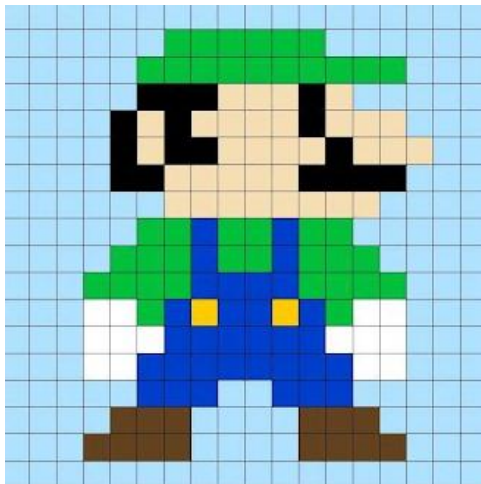
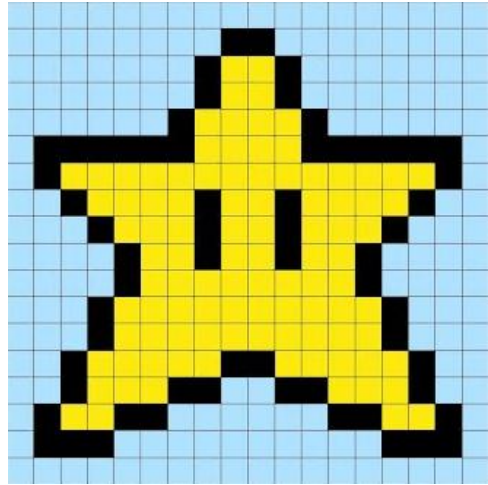
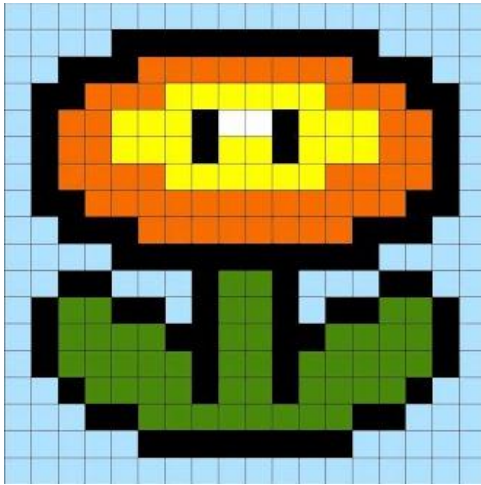
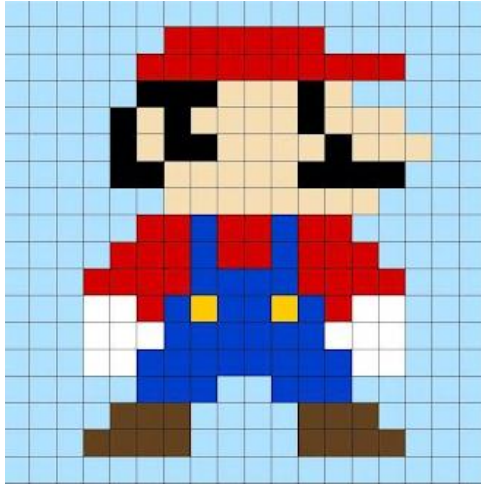


# What path to take?

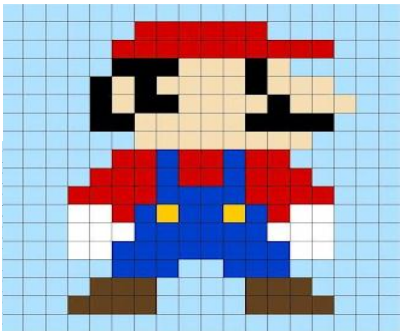
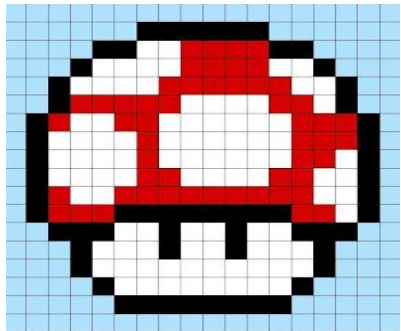
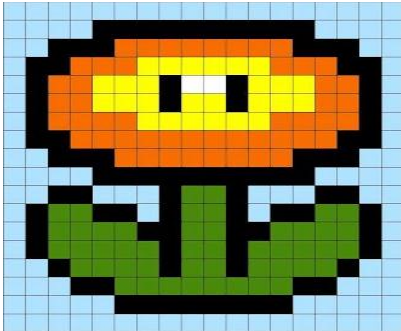
- Mario Circuit 1: The beaten path
- Ghost Valley 2: Moderately rough terrain
- Rainbow Road: Uncharted territory



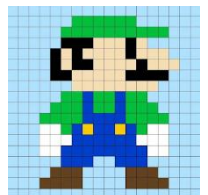
# Data Management Training



# Training: Skills boost



- Identify skills gaps
- Build on existing efforts
- Collaborate
- Relate it back to the mission
- Evaluate & assess
- Revise and refine

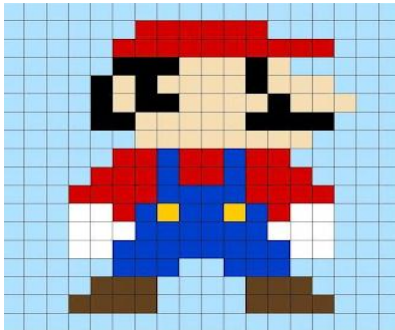


# What level of training do you offer?

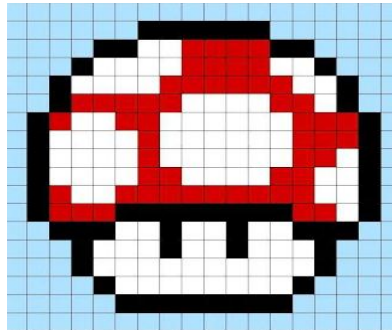
Data management plans & planning



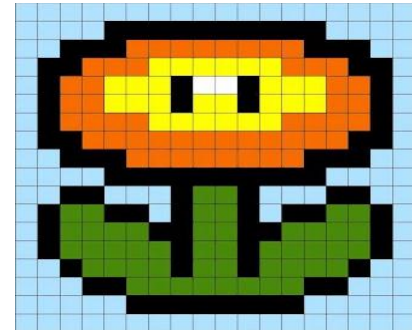
Basic



Moderate



Advanced



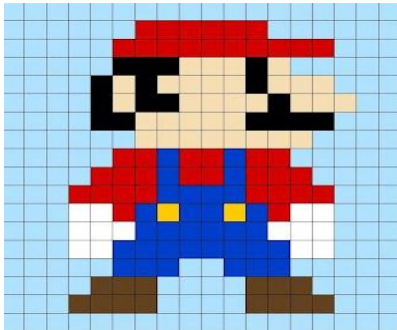


# What level of training do you offer?

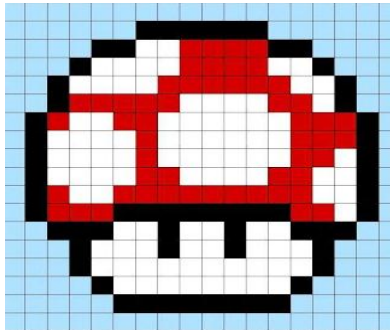
## File organization & naming



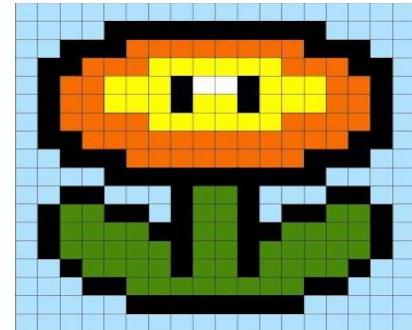
Basic



Moderate



Advanced



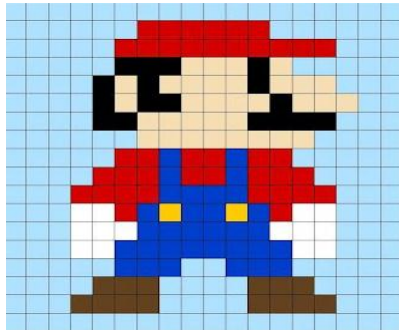


# What level of training do you offer?

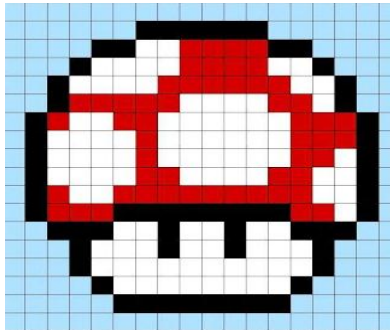
## Backup & storage



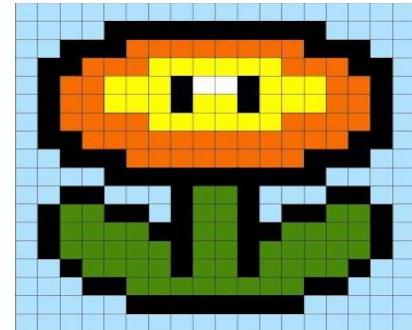
Basic



Moderate



Advanced

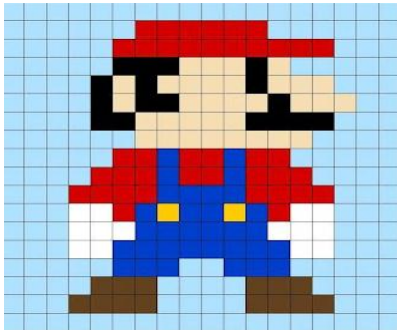


# What level of training do you offer?

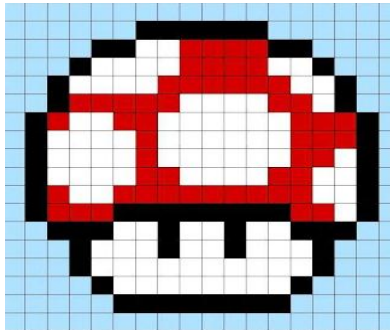
Documentation & metadata



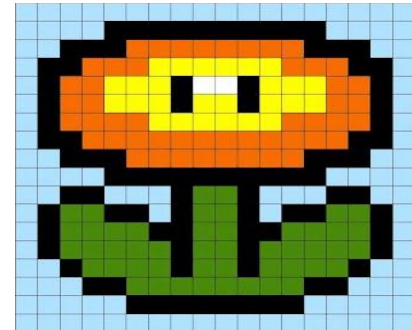
Basic



Moderate



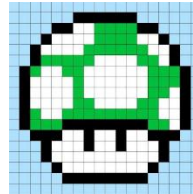
Advanced



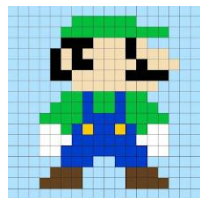
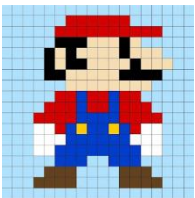
# Training: Various contexts

- Scholarly communication

- Preservation & curation

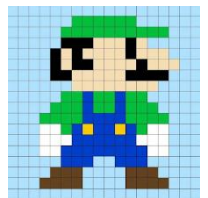
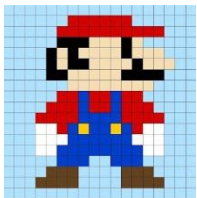


- Responsible Conduct of Research (RCR)



# Training: Strategies

- Make it relevant to expressed/identified needs
- Focus on the practical & concrete
- Demonstrate strategies
- Provide opportunities to practice on examples
- Use real-world datasets
- Target to discipline or research approach

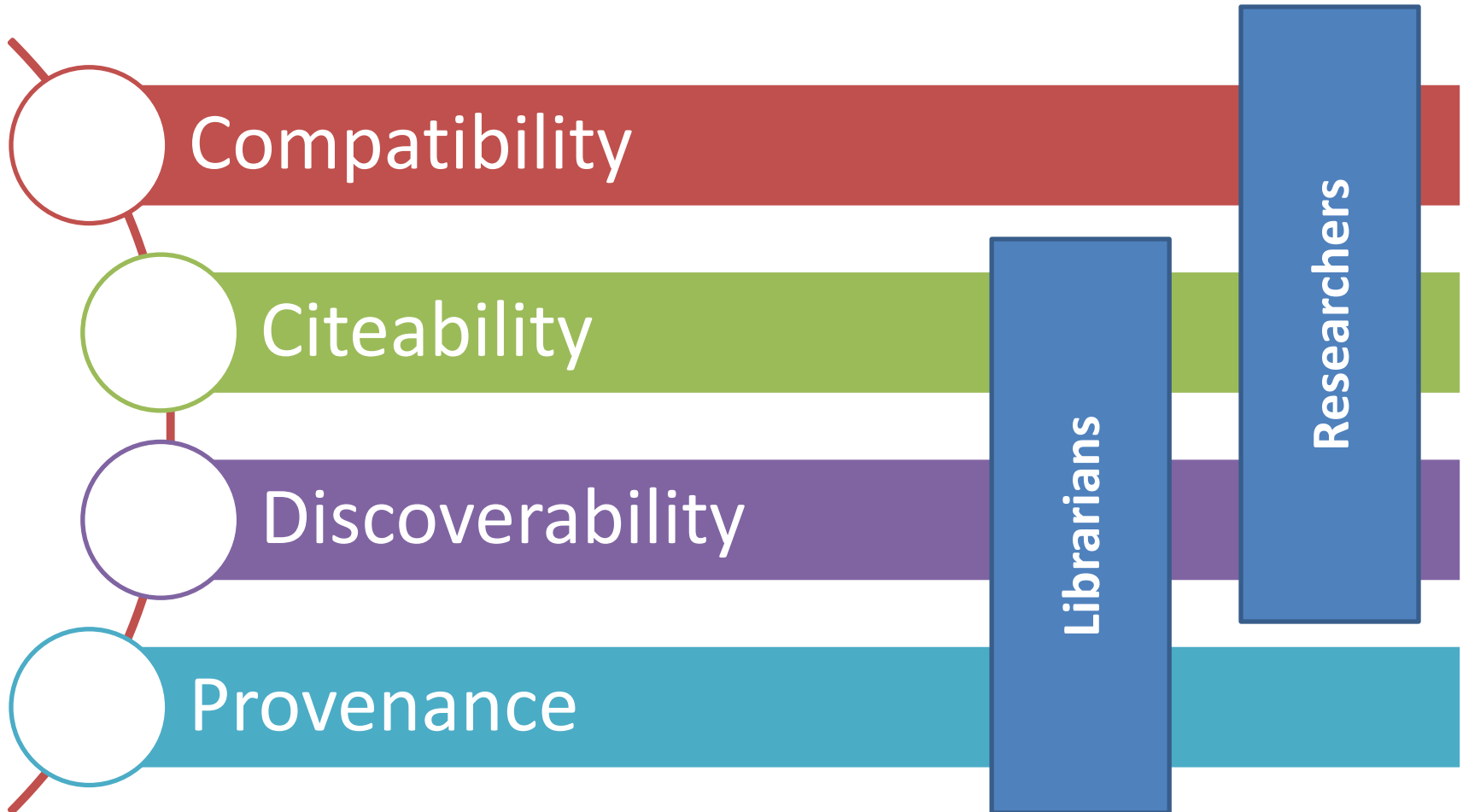




# **METADATA & DOCUMENTATION : AVOIDING PITFALLS**

# The Big Question

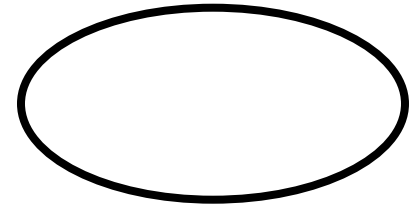
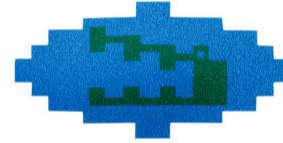
## Metadata: Who Cares?



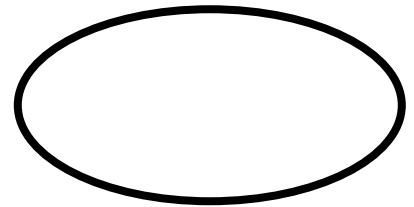


# Things I would rather do than think about scientific metadata:

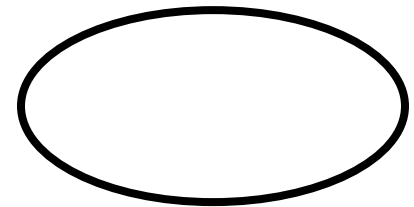
- Fight a pit of crocodiles



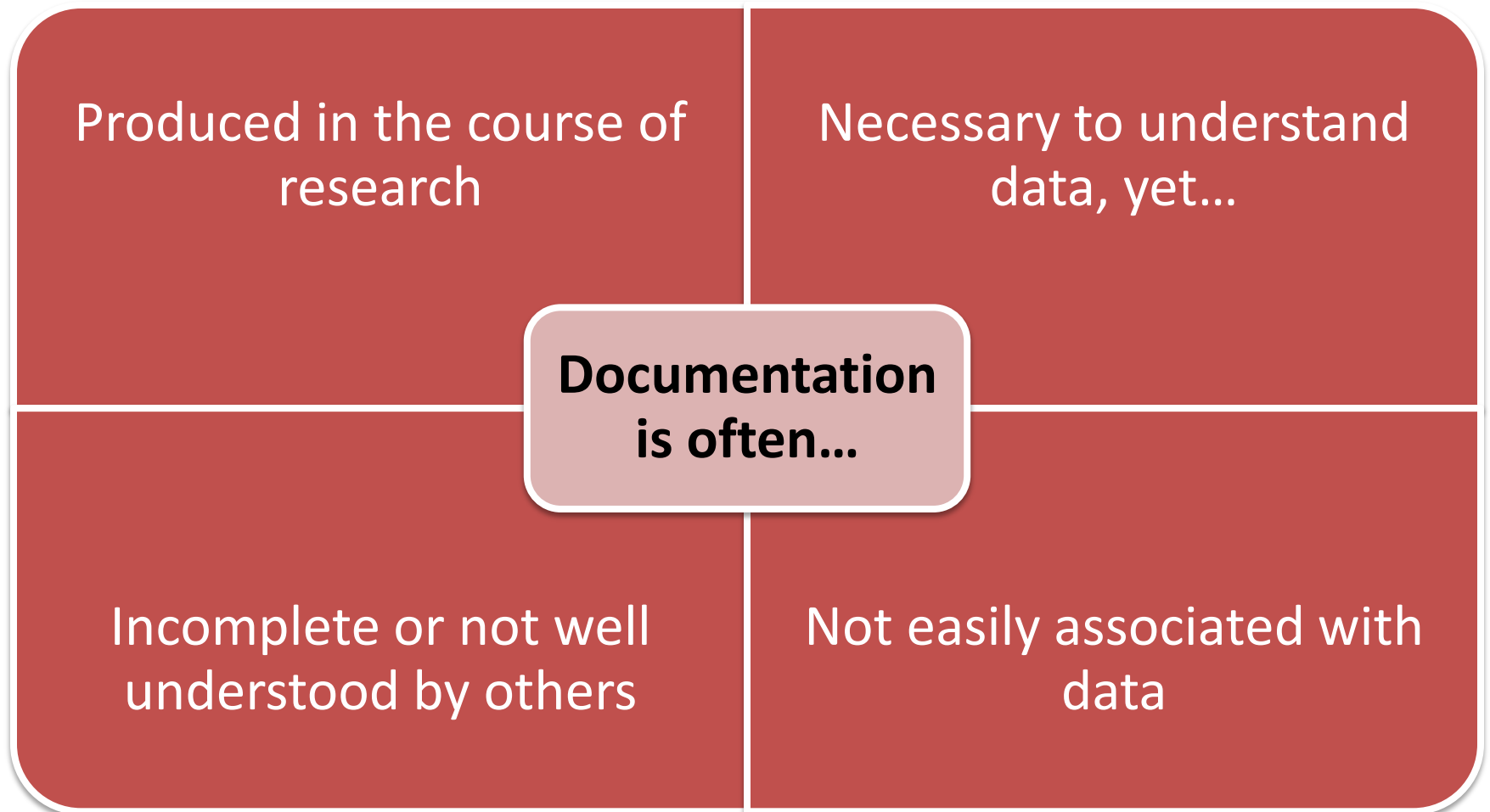
- I'm neutral.



- Nothing! It's very rewarding.



# Metadata & Documentation: Related?



# Documentation

Lab notebooks

Codebooks

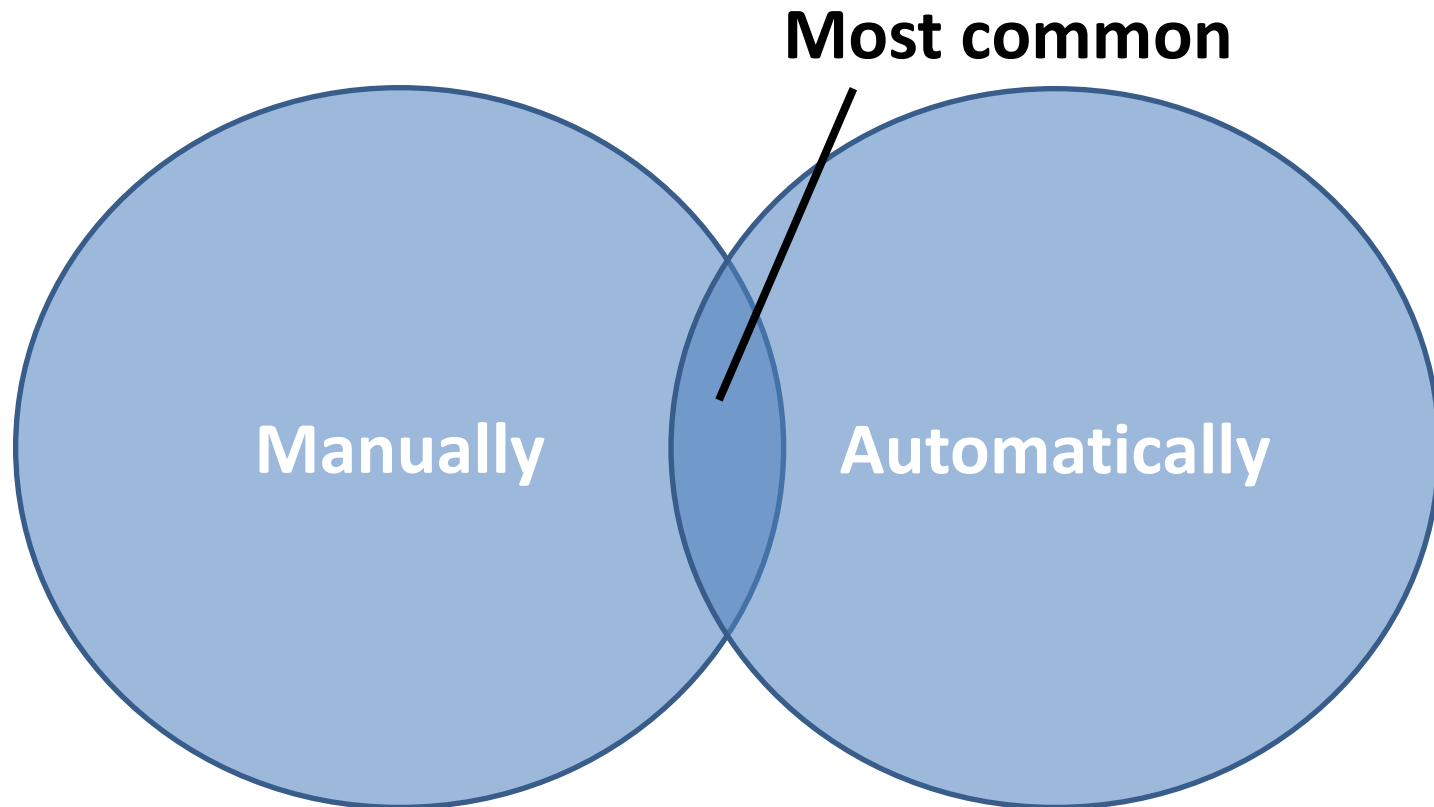
Lab protocols

Grant/research proposals

Associated research articles

# From Documentation to Metadata

Metadata can either be added...



# Metadata : Questions



**What is  
produced  
automatically?**

**Level 1**

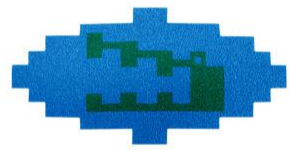
**Level 2**

**What is easily  
transformed into  
relevant  
standards?**

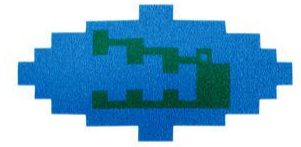


**What information  
is needed to  
open,  
understand,  
& work with the  
data?**

**Level 3**

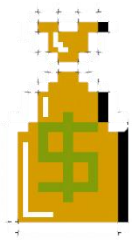


# Metadata : Pitfalls

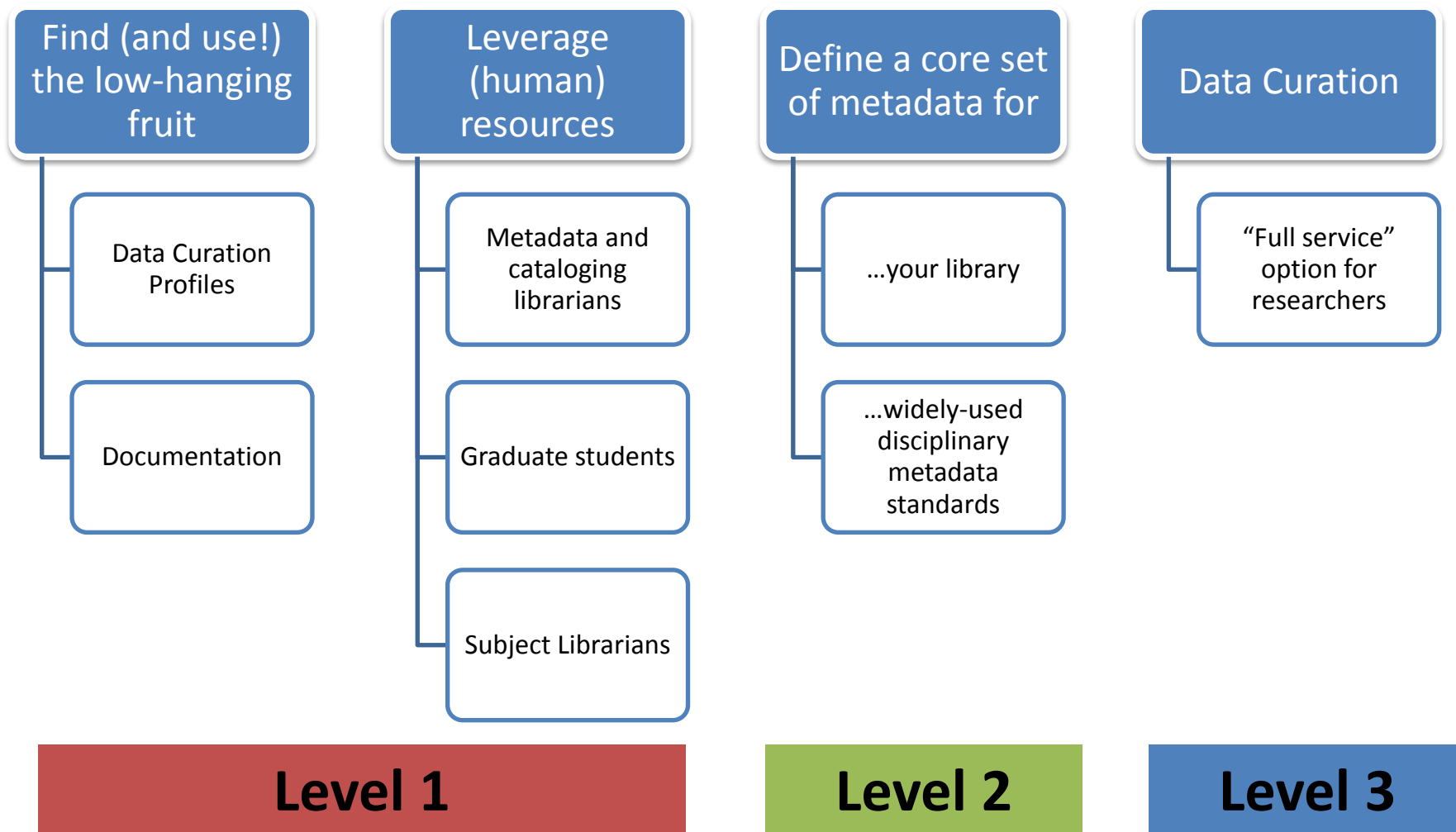


- Rejecting researcher metadata
- Requiring too much of researchers
- Not leveraging existing documentation & expertise
- No easy to understand documentation





# Metadata : Solutions



# Metadata : Requirements

- Data creator name(s)
- Data set title
- File information
  - Format
  - Software required
  - Other technical information
- Methodology
  - How the data was created or collected

## **Vortex II Forecast Data - forecast\_20100615140000Z\_run001**

**Plale, Beth; Brewster, Keith; Mattocks, Craig; Bhangale, Ashish; Withana, Eran C.; Herath, Chathura; Terkhorn, Felix; Chandrasekar, Kavitha**

**URI:** <http://dx.doi.org/10.5967/M0D21VHV>

<http://hdl.handle.net/2022/15157>

**Date:** 2010-07-28

**Date(s) Covered:** 2010-06-15

14:00:00 hours

**Geographic / Spatial Information:** West: -107.1492, East: -95.65079, North: 39.58868, South: 46.41132

Mike Rd Claude TX 79019 USA

**Methodology:** The input data for this forecast includes the following: Rapid Update Cycle (RUC) data downloaded from NOAA with a 13km resolution for forecast date 20100502 at 06Z with data for hourly offsets from 08 to 22. The file format for this input data is grib. The forecast is initialized based on ARPS Data Analysis System (ADAS) Real-time meteorological data assimilation netgrdbas files with CONUS coverage at 10km resolution produced hourly by CAPS at Oklahoma University that uses the netCDF file format. The data is for 20100502 at 13Z.

**File Information:** This particular collection contains namelist.input, cape.zip, radar.zip, precip.zip, surface.zip, updraft\_helicity.zip, vorticity.zip, xsec.zip, and wrfout\_d01\_2010-05-02\_13\_00\_00.nc. namelist is configuration file of WRF. cape is short for Convective Available Potential Energy, a measure of the instability in an air mass. cape.zip is the visualization of cape and contains 24 png files. radar is Mix of radar minimum and radar maximum visualizations. radar.zip represents the mixed results of putting those two radar types together. radar.zip is the visualization of vorticity and contains 28 png files. precip is short for Precipitation, the sum of the rain, snow and hail in given in liquid equivalent depth. precip.zip is the visualization of precip and contains 4 png files. surface is meteorological parameters on the earth's surface, or in a model on the first level above the ground. surface.zip is the visualization of surface and contains 16 png files. updraft\_helicity is the dot product of the vertical velocity and the vertical vorticity. It is presented as a summation over a 3-km depth. updraft\_helicity.zip is the visualization of updraft\_helicity and contains 16 png files. vorticity is the localized rotation of the air. In model plots it is often the vertical component of vorticity, the rotation of the horizontal winds. vorticity.zip is the

**Example of metadata used to describe a data file in IUScholarWorks repository**

**SCORE** 1837170  
**RINGS** 0



# DATA PRESERVATION : THE GOLDEN RING



**SONIC**

x

2

© SEGA

# Data Preservation & Library Roles



- Dedicated data repository
- Sufficient staffing
- Data curation services



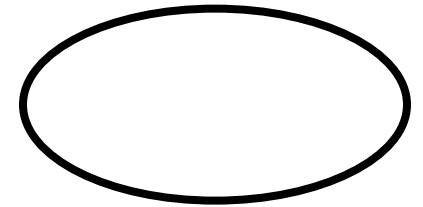
- Leveraged IR & staff
- Consultations & education



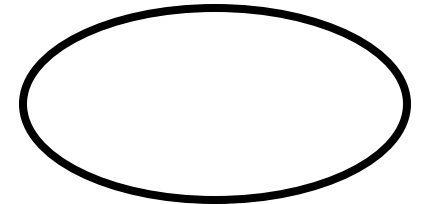
- No IR
- Limited/leveraged staff
- Use of 3<sup>rd</sup> party resources

# Our Library Takes The Role of \_\_\_\_\_ in Data Preservation

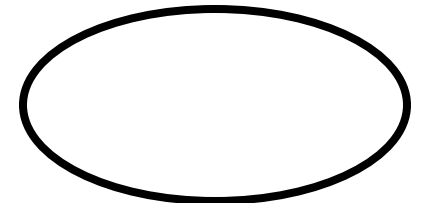
- Lead hedgehog!



- Sidekick



- Elusive genius



# Data Preservation

- How long?
  - Standards vary from discipline to discipline (and grant-to-grant)
- How much?
- What formats?
- At what cost?



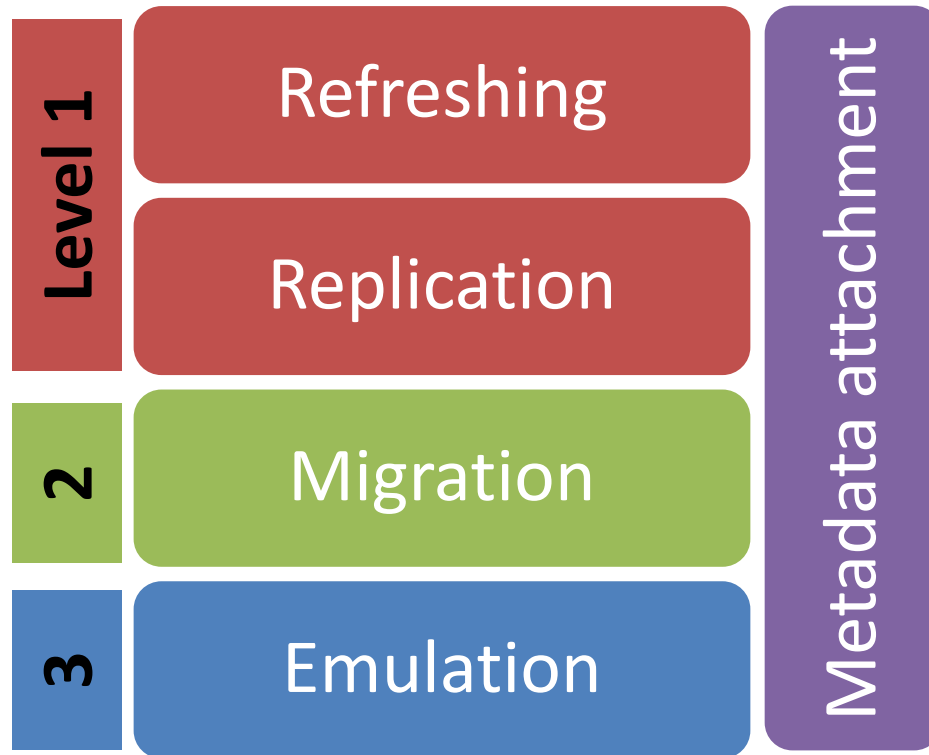
# Data Preservation Challenges

- Was upstream stewardship sufficient?
- Custom and proprietary formats
- Preserving data for emeritus faculty
- Digital data more fragile than analog data in most cases





# Data Preservation



# Data Preservation Options



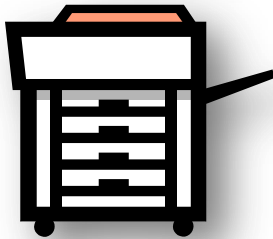
# Data Preservation : Other Opportunities



Partnerships with Archives



Secured physical storage



Replicating your institutions  
digital data





# DATA REPOSITORIES



L=01

## 1. Refer patrons to data repositories

- Find data repositories in [Databib](#)
- Many different flavors: publisher ([Dryad](#)), consortium ([3TU.Datacentrum](#)), instrument ([CHANDRA](#)), sub/disciplinary ([RKMP](#)), country ([Research Data Australia](#)), institutional ([PURR](#)), general purpose ([FigShare](#)), etc.



Have you ever gotten a  
reference question that  
involved a data repository?

YES

NO





## 2. Refactor your existing institutional repository

- 📌 Extending your digital document repository to include dataset submissions
- 📌 For example, [IDEALS](#) at the University of Illinois





## 3. Deploy a dedicated data repository

- 📌 Design and implement a stand-alone institutional data repository
- 📌 For example, the Purdue University Research Repository ([PURR](#))



**Do you have a data repository  
at your institution now?**

**YES**

**NO**





If not, are you planning to  
implement one?

YES

NO



# DATA CITATION

## 1. Instruction



Information literacy outreach



libguides

**PURDUE UNIVERSITY LIBRARIES** *Access. Knowledge. Success.*

Libraries » LibGuides » Citing Data Admin Sign In

### Citing Data

Instructions on citing your use of research datasets

Last Updated: May 25, 2011 | URL: <http://guides.lib.purdue.edu/datacitation> | [Print Guide](#) | [RSS Updates](#) | [Email Alerts](#) | [SHARE](#) | [Facebook](#) | [Twitter](#) | [Google+](#)

Home | [Print Page](#) | Search:  This Guide

#### Examples of Data Citations

Always check your syllabus or author guidelines to see if they contain directions for citing data. Some data distributors will suggest citations that you may use. Most common style guides (e.g., the Chicago Manual of Style) do not give specific instructions for citing data; however, here are three examples from those that do:

**Publication Manual of the American Psychological Association (APA), 6th Edition**

Pew Hispanic Center. (2004). Changing channels and crisscrossing cultures: A survey of Latinos on the news media [Data file and code book]. Retrieved from <http://pewhispanic.org/datasets/>

#### How do I cite data?

When you're writing a research paper, it is necessary to cite your use of sources, typically as footnotes at the bottom of the page or in a bibliography at the end of the paper. It is crucial to provide references for your reader to better understand the context of your research and to give credit for people's work that you've used. As research becomes more data-intensive, it is important to cite your use of datasets in addition to traditional publications such as journal articles, books, and conference proceedings.

Digital datasets come in a wide variety of formats. Some examples include:

- spreadsheets
- interview transcripts
- sensor and instrument readings
- high resolution images
- gene sequences
- software source code
- video recordings

**\* The emerging best practice is to cite data just as you would cite a research article. \***

Most traditional forms of documents are not capable of representing these kinds of data, and so datasets can be published separately in data repositories and other web sites. Whether you produced the data yourself or you're using someone else's data in your research, it is

#### Subject Guide

**Michael Witt**

[Facebook](#) [Twitter](#) [LinkedIn](#)

**Contact Info**  
Stewart Center Room G50  
765-494-8703

**UNIVERSITY OF HAWAII AT MANOA LIBRARY**

[Ask Us](#) | [Research Tools](#) | [Library Catalog](#) | [My Account](#) | [Library](#) » [LibGuides](#) » [Data Management Plans](#) Admin Sign In

### Data Management Plans

Creating a data management plan for access, sharing, and preservation

Last Updated: Nov 19, 2012 | URL: [http://guides.library.manoa.hawaii.edu/data\\_management](http://guides.library.manoa.hawaii.edu/data_management) | [Print Guide](#) | [RSS Updates](#) | [SHARE](#) | [Facebook](#) | [Twitter](#) | [Google+](#)

[What Data?](#) | [Writing the Plan](#) | [Data Preservation](#) | [Citing Data](#) | [Best Practices](#) | [DMP Examples from Manoa](#)

**Citing Data** | [Comments\(0\)](#) | [Print Page](#) | Search:  This Guide

#### Standards for Data Citation

A report from the Board on Research Data and Information of the National Academies' National Research Council, entitled released in November 2012 is available for free download (after registering with an email and password) at [http://www.nap.edu/catalog.php?record\\_id=13564](http://www.nap.edu/catalog.php?record_id=13564). The 239 page report features papers by a wide range of data management experts.

[Comments \(0\)](#)

#### Formats for Data Citation

The International Polar Year Data and Information Service notes that "data citation is a developing practice." At <http://ipdis.org/data/citations.html> the format based on the Chicago Manual of Style 15th edition is: **Authors or group. Date of the release. Title of the data set. Editor or compiler. Place of Publication. Data Publisher. Access date. URI or other distribution method.** "Data publishers (e.g. data centers) have a responsibility to work with data providers and science teams to develop the actual content of the citation."

Examples taken from the IPY page [accessed 2010 Sept 16]

#### Data Citation and Linking

The Digital Curation Centre in the UK has produced a large body of guidelines and advice for data management. This latest report by Alexander Bail and Monica Duke focuses on best practices in data citation, citing data at the most useful level. <http://www.dcc.ac.uk/resources/how-guides/cite-datasets>

Bail, A. & Duke, M. (2011). How to Cite Datasets and Link to Publications. DCC How-to Guides. Edinburgh: Digital Curation Centre. Available online: <http://www.dcc.ac.uk/resources/how-guides>

[Comments \(0\)](#)

#### Citeable Data

The OECD recently (rev. 2010 Feb.) published a white paper about making their datasets citeable. Noting that much of their data are cited without granularity so that the reader cannot easily find the data upon which the analyses and inferences have been made. See <http://dx.doi.org/10.1787/603233448430> for the full report.

DataCite, [www.datacite.org](http://www.datacite.org), is an international collaboration of university libraries and

**Do you include citing data in  
your information literacy  
outreach or libguides?**

**YES**

**NO**



# DATA CITATION

## 2. Outreach and Advocacy

- 🌟 Brochure
- 🌟 Writing Lab
- 🌟 Press and publishers



# DATA CITATION

## 3. Practice what you preach

- ☒ Cite your own data
- ☒ Suggest citation for data in your libraries
- ☒ Include in supporting documentation
- ☒ COiNS, embedded RDF, BibTex, etc.
- ☒ Show publications that cite datasets
- ☒ DataCite: minting DOIs for datasets



[www.DataCite.org](http://www.DataCite.org)

# Resources

- Tenopir, C., Birch, B., & Allard, S. (2012). Academic Libraries and Research Data Services: Current Practices and Plans for the Future. [http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/Tenopir\\_Birch\\_Allard.pdf](http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/Tenopir_Birch_Allard.pdf)
- Bailey, C. (2012). Research Data Curation Bibliography: <http://digital-scholarship.org/rdcb/rdcb.htm>
- Walters, T., Skinner, K., & Association of Research Libraries. (2011). *New Roles for New Times: Digital Curation for Preservation*. [http://www.arl.org/bm~doc/nrnt\\_digital\\_curation17mar11.pdf](http://www.arl.org/bm~doc/nrnt_digital_curation17mar11.pdf)
- Association of Research Libraries. *To Stand the Test of Time: Long-Term Stewardship of Digital Data Sets in Science and Engineering*. <http://www.arl.org/bm~doc/digdatarpt.pdf>
- Jahnke, L., Asher A., & Keralis, S. (2012). *The Problem of Data*. Council on Library and Information Resources. <http://www.clir.org/pubs/reports/pub154>
- Borgman, C. L. (2012). The Conundrum of Sharing Research Data. *Journal of the American Society for Information Science and Technology* 63(6), 1059-1078. <http://ssrn.com/abstract=1869155>